

Are On-Line Personae Really Unlinkable?

What the US government sees...



What anybody can see...



One-to-one linkage



Attribute	Value	Linkage	Attribute	Value
Age	35	→	Age	35
Gender	Male	→	Gender	Male
City	London	→	City	London
Country	UK	→	Country	UK
Occupation	Software Engineer	→	Occupation	Software Engineer
Interests	Technology, Music	→	Interests	Technology, Music
Education	University of Cambridge	→	Education	University of Cambridge
Marital Status	Single	→	Marital Status	Single
Religion	Christianity	→	Religion	Christianity
Political Affiliation	Conservative Party	→	Political Affiliation	Conservative Party
Travel History	USA, Japan, Australia	→	Travel History	USA, Japan, Australia
Language	English	→	Language	English
Favorite Color	Blue	→	Favorite Color	Blue
Favorite Food	Pasta	→	Favorite Food	Pasta
Favorite Music	Rock	→	Favorite Music	Rock
Favorite Sport	Football	→	Favorite Sport	Football
Favorite TV Show	Game of Thrones	→	Favorite TV Show	Game of Thrones
Favorite Book	The Hobbit	→	Favorite Book	The Hobbit
Favorite Movie	The Lord of the Rings	→	Favorite Movie	The Lord of the Rings
Favorite Animal	Cats	→	Favorite Animal	Cats
Favorite Flower	Roses	→	Favorite Flower	Roses
Favorite Color	Blue	→	Favorite Color	Blue
Favorite Food	Pasta	→	Favorite Food	Pasta
Favorite Music	Rock	→	Favorite Music	Rock
Favorite Sport	Football	→	Favorite Sport	Football
Favorite TV Show	Game of Thrones	→	Favorite TV Show	Game of Thrones
Favorite Book	The Hobbit	→	Favorite Book	The Hobbit
Favorite Movie	The Lord of the Rings	→	Favorite Movie	The Lord of the Rings
Favorite Animal	Cats	→	Favorite Animal	Cats
Favorite Flower	Roses	→	Favorite Flower	Roses

Record linkage:
common attributes, rare values

What the US government sees....

TOP SECRET//SI//ORCON//NOFORN

Gmail facebook Hotmail^{msn} Google^{apple} skype paltalk.com YouTube AOL mail

 (TS//SI//NF) PRISM Collection Details 

Current Providers

- Microsoft (Hotmail, etc.)
- Google
- Yahoo!
- Facebook
- PaITalk
- YouTube
- Skype
- AOL
- Apple

What Will You Receive in Collection (Surveillance and Stored Comms)?
It varies by provider. In general:

- E-mail
- Chat – video, voice
- Videos
- Photos
- Stored data
- VoIP
- File transfers
- Video Conferencing
- Notifications of target activity – logins, etc.
- Online Social Networking details
- **Special Requests**

Complete list and details on PRISM web page:
Go PRISMFAA

TOP SECRET//SI//ORCON//NOFORN

What anybody can see....

meilof's Music Profile - Users at Last.fm - Mozilla Firefox

Firefox Your prezis | Prezi meilof's Music Profile - User... +

www.last.fm/user/meilof

last.fm Music Events Charts Community Originals

Inbox | Logout meilof

Come work with us! Last.fm is hiring » English | Paint it Black has moved | Help Music search

Gratis Samsung Galaxy S4
Nú 6 maanden van € 59 voor € 29,50 p.mnd.
Onbeperkt bellen en 2GB data
Klik hier

DE ZOMER VAN VODAFONE

Gratis iPhone 5
Nú 6 maanden van € 64 voor € 32 p.mnd.
Toestel is van jou
Klik hier

Met 2 GB data

vodafone

About Me
And it's getting strange in here
Yeah, it gets stranger every year
More news from nowhere
More news from nowhere

Recent Activity
You tagged Farid Mammadov - Hold Me, Eyþór Ingi Gunnlaugsson - Eg á líf (Eurovision 2013 - Iceland), Emilia and 11 other items with half toontje hoger. May 2013

meilof Library Friends Tracks Albums Charts Neighbours Events More...

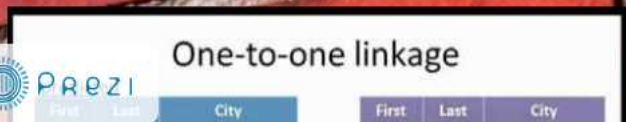
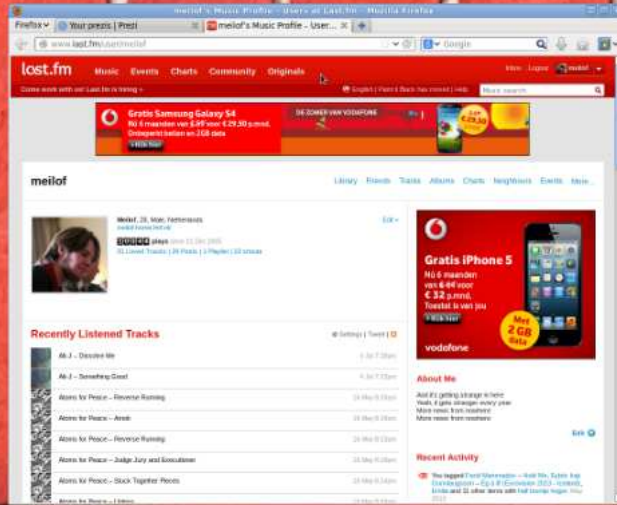
Meilof, 28, Male, Netherlands
meilof.home.tmf.nl Edit »

10344 plays since 21 Dec 2005
31 Loved Tracks | 26 Posts | 1 Playlist | 22 shouts

Recently Listened Tracks Settings | Tweet |

	Alt-J - Dissolve Me	4 Jul 7:28pm
	Alt-J - Something Good	4 Jul 7:23pm
	Atoms for Peace - Reverse Running	16 May 8:33pm
	Atoms for Peace - Amok	16 May 8:26pm
	Atoms for Peace - Reverse Running	16 May 8:23pm
	Atoms for Peace - Judge Jury and Executioner	16 May 8:19pm
	Atoms for Peace - Stuck Together Pieces	16 May 8:14pm
	Atoms for Peace - Unless	16 May 8:09pm

What anybody can see....



Record linkage:

What anybody can see....



One-to-one linkage

predicted			actual		
First Name	Last Name	City	First Name	Last Name	City
john	smith	nashville	john	smith	nashville
bill	clinton	washington dc	bill	clinton	washington dc
hillary	clinton	washington dc	william	clinton	washington dc

Record linkage:

common attributes, rare values




How big is the on-line privacy problem?


multiple databases

 social networking site


First Name	M/F	DOB	Favorite Movie	Favorite Band	#L	Lives in	Nickname
Bruno	M	17/5/1988	Cidade de Deus (2002)	Incubus	3	Porte Alegre	bruno31
Deborah	F	2/9/1962	Saving Private Ryan (1998)	Lady Gaga	1	Seattle	DEBORrH
...

 music discussion platform

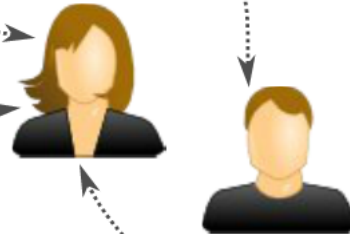
Nickname	Country	M/F	Favorite Band
748	UK	F	Cher Lloyd
DEBORrH	US	F	The Decemberists
...

 movie reviewing website

Nickname	First name	Country	Age	Favorite Movie
BfZhcrinm	Jenny	GER	21	Austin Powers (1997)
bruno	Bruno	BRA	23	Cidade de Deus (2002)
...

 professional networking site

First Name	Surname	M/F	#L	Lives in
Bruno	Branco	M	3	São Paulo
Deborah	Taylor	F	2	Las Vegas
...



common attributes & values



List of most popular given names - Wikipedia, the free encyclopedia - Mozilla Firefox

en.wikipedia.org/wiki/List_of_most_popular_given_names#North_and_South_America

North and South America

Male names

Region (year)	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10
Buenos Aires, Argentina (2012) ^[100]	Juan	Benjamin	Santiago	Thiago	Lucas	Joaquin	Santino	Lautaro	Ian	Mateo
Aruba (2005) ^[101]	Daniel	Dylan/Dyllan	Kevin/Keven	NA	NA	NA	NA	NA	NA	NA
Brazil (2012) ^[102]	Miguel	Arthur	Davi	Gabriel	Lucas	Matheus	Pedro			
Canada (2012, unofficial list) ^[103]	William	Jacob	Liam	Nathan	Noah	Ethan	Lucas/L			
Canada, Alberta (2012) ^[104]	Liam	Ethan	Jacob	Logan	Mason	Benjamin	Lucas			
Canada, British Columbia (2012) ^[105]	Ethan	Liam	Lucas	Mason	Logan	Noah	Alexand			
Canada, Manitoba (2012)	Liam	Mason	Carter	Noah	Logan	Lucas	William			

lost.fm

BEST OF 2011

Top Artists | New Discoveries | Year in Music

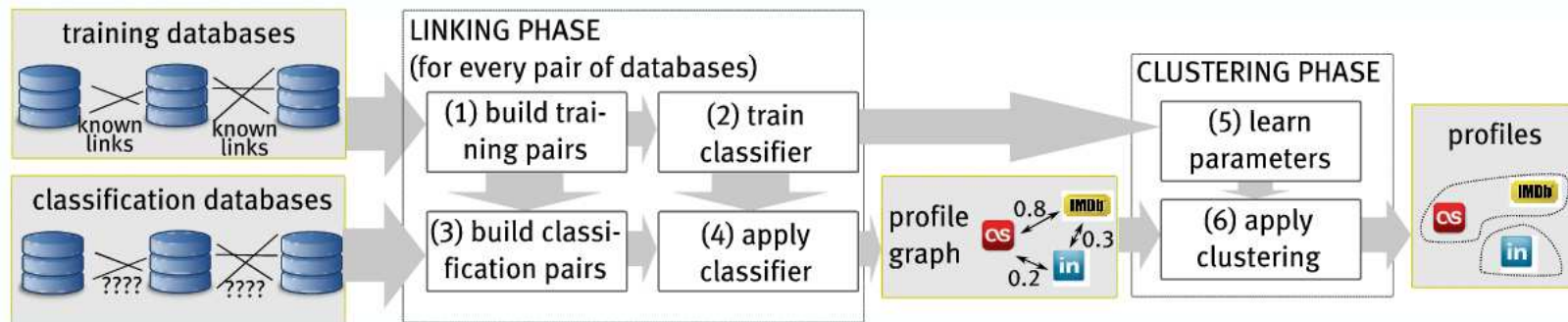
Most listened to artists of 2011

All Music in United Kingdom

Discover 2011 in scrobbles

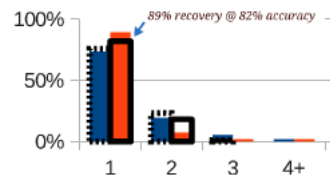
```
*train.csv (/scratch/experiments/files/output_generator) - ge
File Edit View Search Tools Documents Help
Open Save Undo
*train.csv x
"RUTH", "RUTH", "RUTH", "female", "female", "female", "2", "2", "The Shawshank Redemption (1994)", "Star Wars:
Episode V - The Empire Strikes Back (1980)", "Bon Iver", "Bon Iver", "2", "1", "Colorado Springs", "Colorado
Springs", "us", "us", "us", "us", "RUTH", "RUTHImZaR", "RUTHImZaR", "22", "22"
"Seougia", "Seougia", "Seougia", "female", "female", "female", "21", "21", "The Avengers (2012)", "The Avengers
(2012)", "Britney Spears", "Britney
Spears", "4", "4", "London", "London", "uk", "uk", "uk", "uk", "yY", "yY", "yY", "6", "6"
"Rafaela", "Rafaela", "Rafaela", "female", "female", "female", "3", "3", "Orfeu Negro (1959)", "Orfeu Negro
(1959)", "Avril Lavigne", "Radiohead", "2", "0", "Porto
Alegre", "Londrina", "brazil", "brazil", "brazil", "brazil", "McR", "Rafdhla", "McR", "39", "39"
Plain Text Tab Width: 8 Ln 1, Col 1 INS
```


experimental numbers

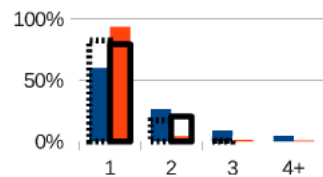


experimental numbers

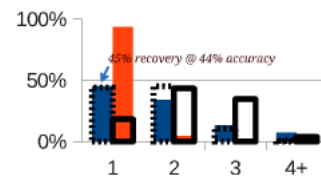
[A] ED/EC: bline,dense,np



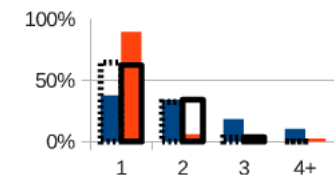
[B] ED/EC: bline,dense,np



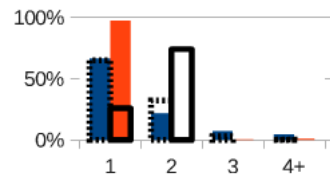
[C] ED/EC: bline,dense,pt



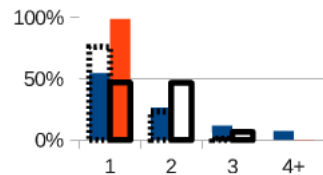
[D] ED/EC: bline,sparse,pt



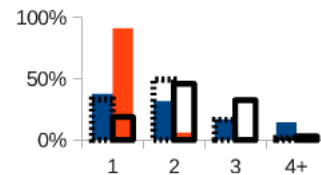
[E] ED/EC: large,dense,np



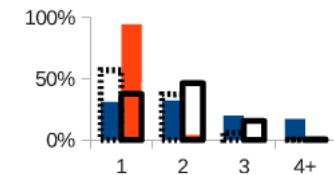
[F] ED/EC: large,sparse,np



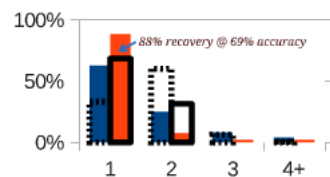
[G] ED/EC: large,dense,pt



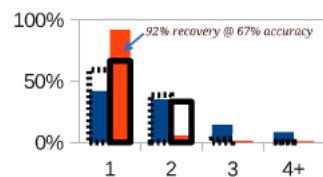
[H] ED/EC: large,sparse,pt



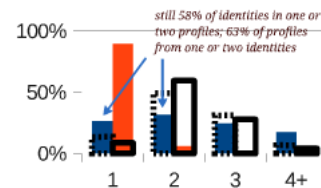
[I] ED/EC: low-o,dense,np



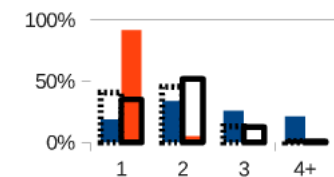
[J] ED/EC: low-o,sparse,np



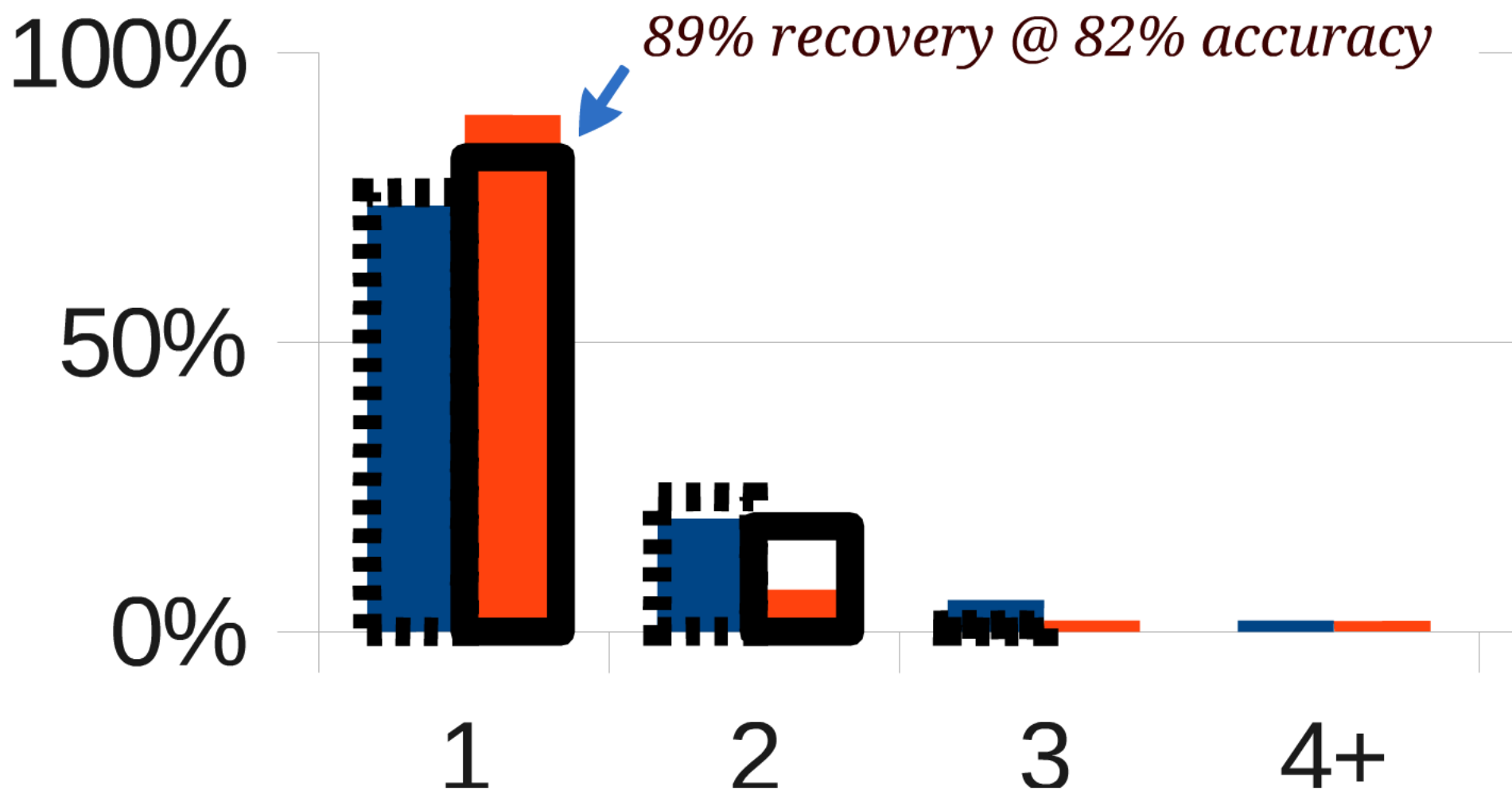
[K] ED/EC: low-o,dense,pt



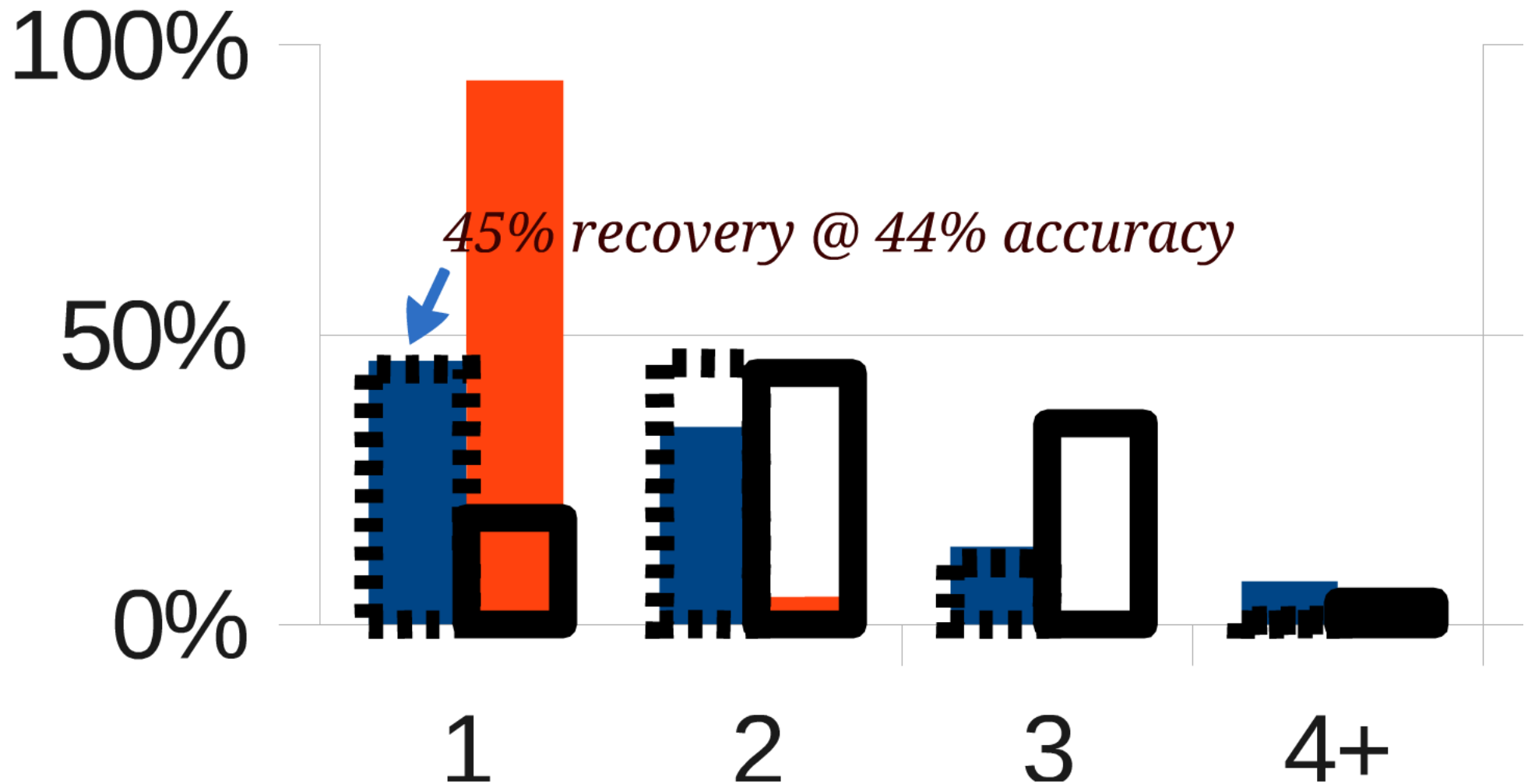
[L] ED/EC: low-o,sparse,pt



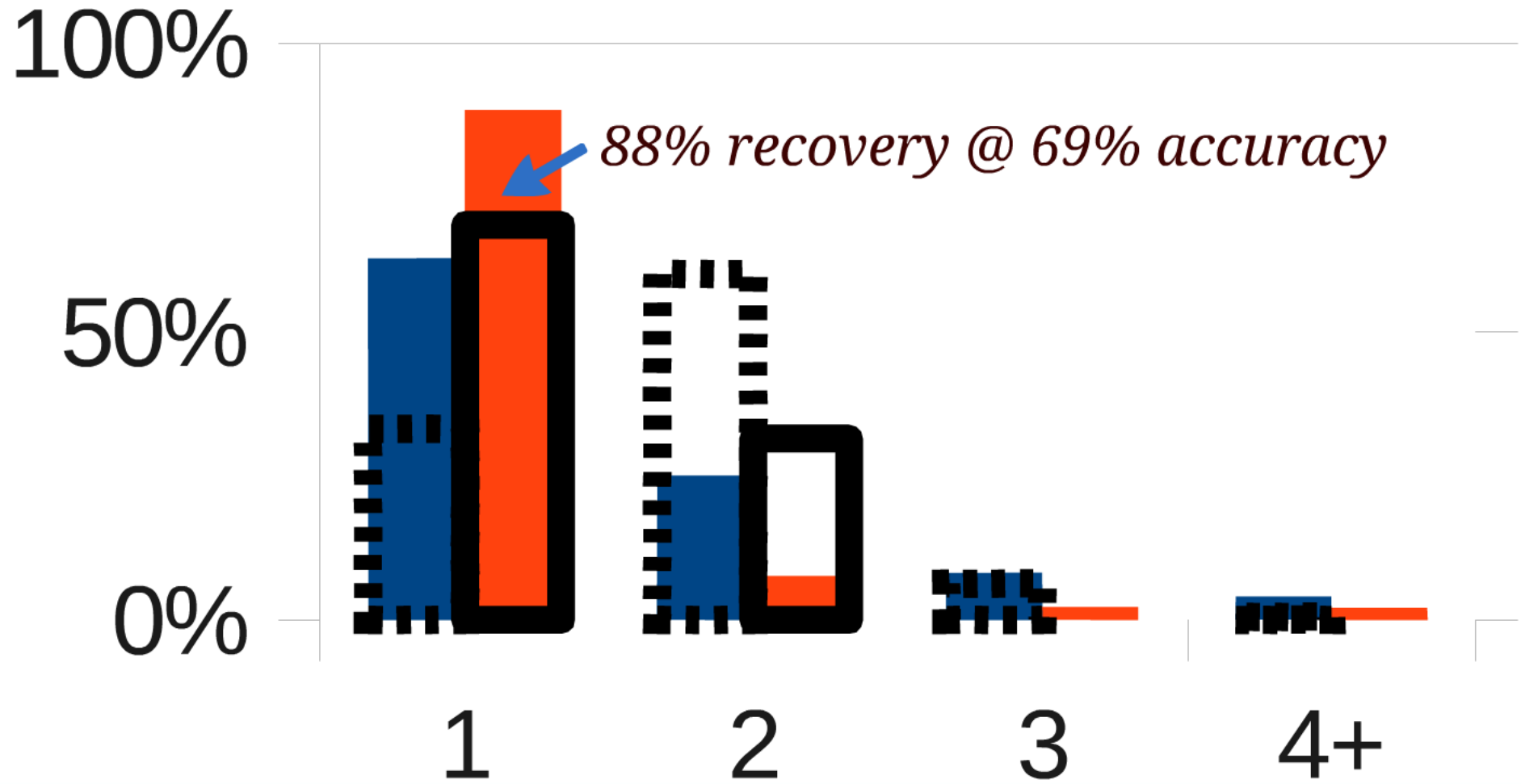
[A] ED/EC: bline,dense,np



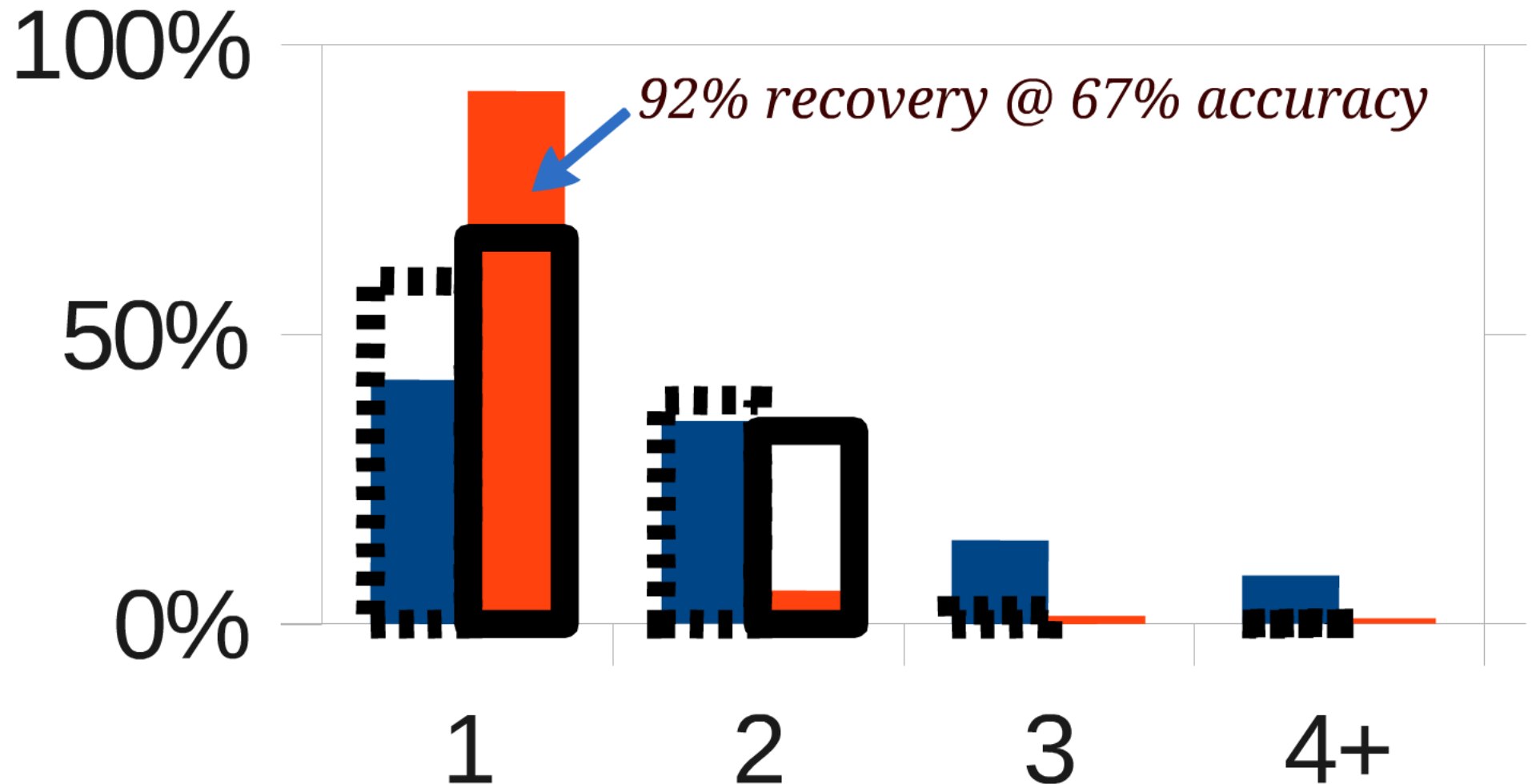
[C] ED/EC: bline,dense,pt



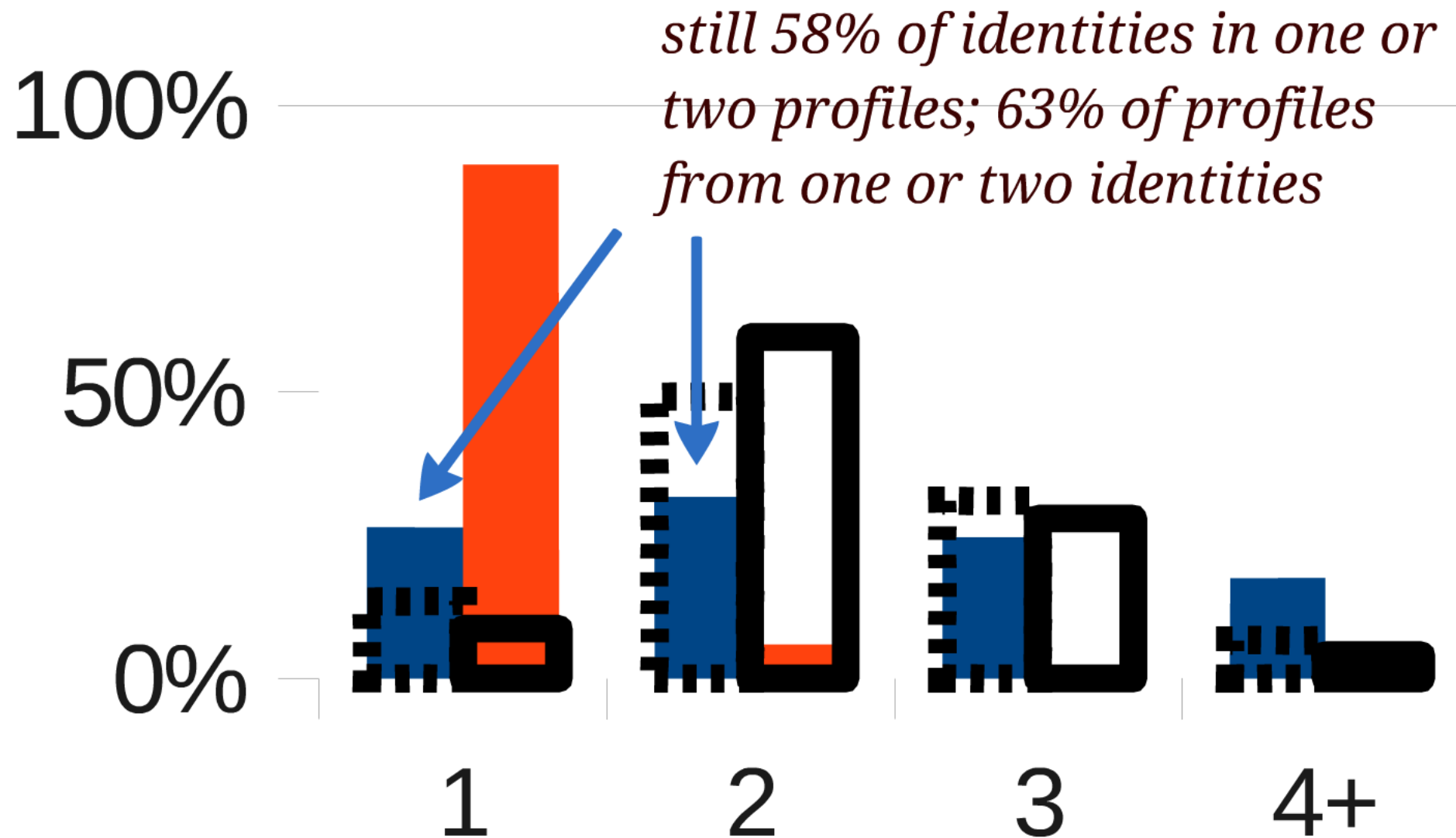
[I] ED/EC: low-o,dense,np



[J] ED/EC: low-o,sparse,np

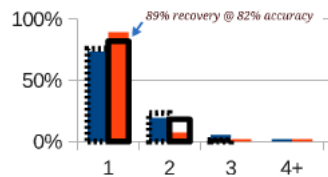


[K] ED/EC: low-o, dense, pt

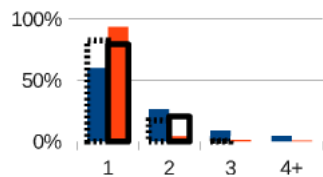


experimental numbers

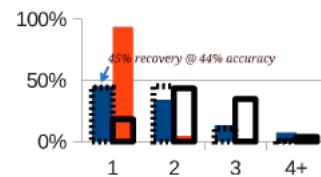
[A] ED/EC: bline,dense,np



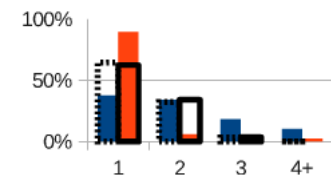
[B] ED/EC: bline,dense,np



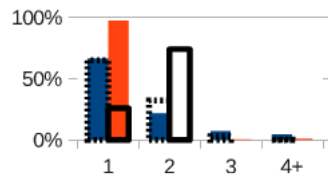
[C] ED/EC: bline,dense,pt



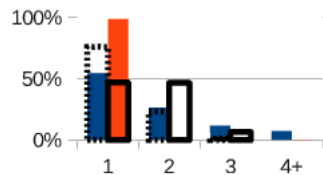
[D] ED/EC: bline,sparse,pt



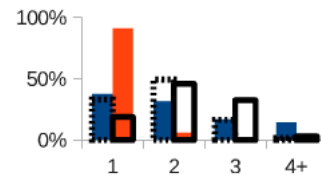
[E] ED/EC: large,dense,np



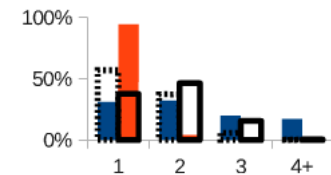
[F] ED/EC: large,sparse,np



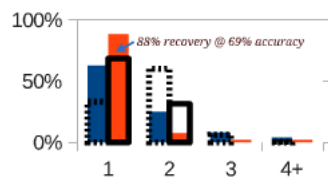
[G] ED/EC: large,dense,pt



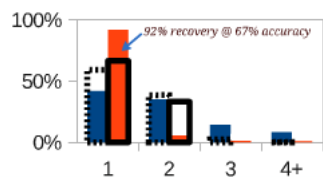
[H] ED/EC: large,sparse,pt



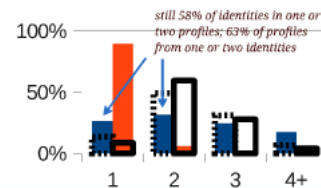
[I] ED/EC: low-o,dense,np



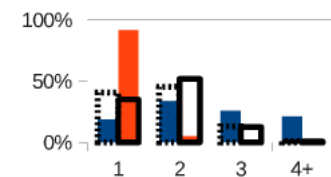
[J] ED/EC: low-o,sparse,np



[K] ED/EC: low-o,dense,pt



[L] ED/EC: low-o,sparse,pt

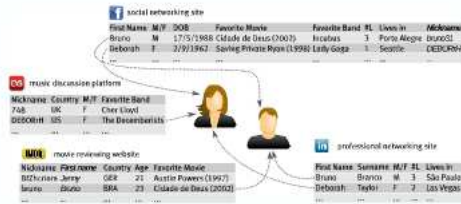


experimental numbers

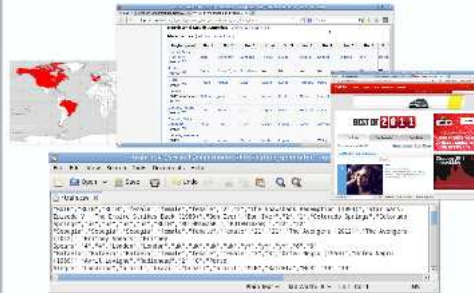
Table 1. Average precision (“p”)/recall (“r”)/f-measure (“f”) results of our experiments after linking phase (“PW”); community detection (“CD”); and threshold transitive closure (“TC”). Boldfaced f-measures indicates significantly best result(s). “Pt” means perturbation.

	Dense, no pt			Sparse, no pt			Dense, pt			Sparse, pt			
	p	r	f	p	r	f	p	r	f	p	r	f	
Baseline	PW	0.37	0.83	0.51	0.38	0.83	0.52	0.24	0.68	0.36	0.25	0.68	0.37
	CD	0.75	0.84	0.79	0.50	0.82	0.62	0.48	0.58	0.52	0.35	0.60	0.44
	TC	0.49	0.91	0.63	0.62	0.83	0.70	0.57	0.30	0.39	0.28	0.60	0.38
Large	PW	0.23	0.83	0.36	0.24	0.83	0.37	0.14	0.68	0.23	0.14	0.68	0.23
	CD	0.64	0.76	0.69	0.42	0.74	0.53	0.35	0.48	0.40	0.26	0.51	0.34
	TC	0.52	0.50	0.51	0.91	0.42	0.57	0.20	0.32	0.24	0.41	0.25	0.31
Low Overlap	PW	0.26	0.67	0.38	0.26	0.67	0.37	0.14	0.43	0.21	0.14	0.44	0.21
	CD	0.50	0.44	0.47	0.32	0.51	0.39	0.20	0.23	0.22	0.15	0.27	0.20
	TC	0.45	0.85	0.58	0.58	0.71	0.64	0.34	0.22	0.26	0.38	0.22	0.28

multiple databases



common attributes & values



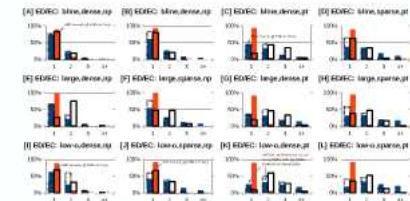
experimental numbers



*Conclusion:
non-identifying data can often
be linked & clustering helps*

*Future work:
overcoming our limitations:
approaches & experiments & insight*

experimental numbers



experimental numbers

Table 1. Average precision (P), recall (R), F-measure (F) results of our experiments after linking phase ("PW"), community detection ("CD"), and threshold transitive closure ("TC"). Boldfaced F-measures indicates significantly best results(s). "Pc" means perturbation.

		Dense, no pt		Sparse, no pt		Dense, pt		Sparse, pt		
		P	R	F	P	R	F	P	R	
Baseline	PW	0.37	0.83	0.51	0.38	0.83	0.52	0.24	0.68	0.36
	CD	0.75	0.84	0.79	0.50	0.82	0.62	0.40	0.58	0.52
	TC	0.49	0.91	0.63	0.62	0.85	0.70	0.57	0.30	0.30
Large	PW	0.23	0.83	0.36	0.24	0.83	0.37	0.14	0.68	0.23
	CD	0.64	0.76	0.69	0.42	0.74	0.53	0.35	0.48	0.40
	TC	0.57	0.50	0.51	0.91	0.42	0.57	0.20	0.32	0.24
Low Overlap	PW	0.26	0.67	0.38	0.26	0.67	0.37	0.14	0.43	0.21
	CD	0.50	0.44	0.47	0.22	0.51	0.39	0.20	0.23	0.22
	TC	0.45	0.85	0.58	0.58	0.71	0.64	0.34	0.22	0.26